

REMARKS

Applicants respectfully request that the above-identified application be re-examined.

The September 25, 2003, Office Action ("Office Action") rejected all of the claims in this application (1-20) as fully anticipated by the teachings of U.S. Patent No. 6,418,433 (Chakrabarti et al.). While applicants believe that this rejection was clearly in error, in order to advance the prosecution of this application, minor amendments have been made to the claims directed to making them more particularly point out and distinctly claim the subject matter applicants regard as their invention.

Prior to discussing in detail why applicants believe that all of the claims in this application are allowable, a brief description of applicants' invention and a brief description of the teachings of the cited and applied reference are provided. The following discussion of applicants' invention and the cited and applied reference are not provided to define the scope or interpretation of any of the claims of this application. Instead, these discussions are provided to help the United States Patent and Trademark Office better appreciate important claim distinctions discussed thereafter.

The Invention

Applicants' invention is directed to an improved way of retrieving information pertaining to documents stored in a computer network. More specifically, the invention employs a probabilistic model to determine the likelihood that the document has changed since it was last accessed and, thus, to determine whether a document should be accessed during a current Web crawl. Preferably, the accuracy of the probabilistic model is continuously improved by training internal probability distributions to reflect the actual change rate pattern of the document to be accessed.

In one form, the invention is directed to a computer-implemented method for selectively accessing a document during a current crawl of a server computer, the document being identified by a document address specification and having been retrieved during a previous crawl. The method comprises determining whether to access the document during the current crawl with the aid of a probabilistic model that is based on the probability that the document has changed since the previous crawl. The method further comprises accessing the document if the determination produces an instruction indicative that the document at the document address specification should be accessed during the current crawl.

In another form, the invention is directed to a computer-readable medium having computer-executable instructions for retrieving one document in a plurality of documents from a

LAW OFFICES OF
CHRISTENSEN O'CONNOR JOHNSON KINDNESS^{PC}
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100

remote server. When executed, the instructions: (i) maintain historical information associated with changes to the one document; (ii) initiate a crawl procedure for retrieving particular documents in the plurality of documents; and (iii) determine whether to access the one document from the remote server based on a probabilistic analysis of the historical information associated with the changes to the one document.

U.S. Patent No. 6,418,433 (Chakrabarti et al.)

Chakrabarti et al. purportedly discloses a system and method for focused Web crawling. Chakrabarti et al.'s focused Web crawler learns to recognize Web pages that are relevant to the interest of one or more users from a set of examples provided by the users. Chakrabarti et al.'s focused Web Crawler explores the Web starting from the example set using statistics collected from the samples and other analysis of the link graph of a growing crawl database to guide itself toward relevant valuable resources and away from irrelevant and/or low quality material. The Web crawler allegedly builds a comprehensive topic-specific library for the benefit of specific users. In summary, Chakrabarti et al. is directed to locating relevant documents. Chakrabarti et al. is not directed to determining whether a document should be accessed during a current crawl based on the probability that the document has changed since a prior crawl.

While Chakrabarti et al. purportedly does teach revisiting documents, the basis for revisiting the documents is not based on a probabilistic model that the document has changed since a previous crawl. Chakrabarti et al.'s basis for revisiting a document is based on the relevant priority of the document to the subject matter, not the probability that the document has been changed since the previous crawl. See Col. 10, lines 32-34, which states "the more relevant the page, the higher the priority for revisitation to check for subsequent changes." Lines 10-15 of Col. 10 of Chakrabarti et al. state that a page is classified as "good" in terms of relevancy using a topic analyzer.

Claims 1-20

As noted above, the Office Action rejected Claims 1-20 under 35 U.S.C. § 102(e) as being fully anticipated by the teachings of Chakrabarti et al. The Office Action asserts that Chakrabarti et al. suggests each and every element of these claims. Applicants respectfully disagree. As described in more detail below, Chakrabarti et al. fails to disclose or suggest elements of both independent and dependent claims of this application.

As amended, independent Claim 1 reads as follows:

LAW OFFICES OF
CHRISTENSEN O'CONNOR JOHNSON KINDNESSSM
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100

1. A computer-implemented method for selectively accessing a document during a current crawl of a server computer, the document being identified by a document address specification, the document having been retrieved during a previous crawl, the method comprising:

determining whether to access the document during the current crawl with the aid of a probabilistic model that is based on the probability that the document has changed since the previous crawl; and

accessing the document if the determination produces an instruction indicative that the document at the document address specification should be accessed during the current crawl.

As noted above, Chakrabarti et al. clearly does not disclose determining whether to access a document during a current crawl with the aid of a probabilistic model that is based on the probability that the document has changed since a previous crawl. Thus, Chakrabarti et al. also does not disclose accessing the document if the determination produces an instruction indicative that the document address specification should be accessed during the current crawl. Consequently, applicants respectfully submit that Claim 1, particularly as amended, is not anticipated by Chakrabarti et al. and, thus, is allowable.

Applicants also submit that Claims 2-9, all of which depend directly or indirectly from Claim 1, are allowable for the same reasons that Claim 1 is allowable.

Applicants further submit that Claims 2-9 are allowable for additional reasons. For example, Claim 2, which depends from Claim 1, recites that determining whether to access a document with the aid of a probabilistic model comprises computing a probability that the document has changed since the document was created during the previous crawl. Chakrabarti et al. teaches no such computation. As a result, applicants respectfully submit that Claim 2 is allowable for reasons in addition to the reasons why Claim 1 is allowable.

Claim 3, which depends upon Claim 2, recites that computing a probability that the document has changed comprises: selecting an active probability indicative of a proportion of documents in a plurality of documents that are changing at various rates, the plurality of documents including the document; training the active probability to reflect experience with the document during a plurality of previous crawls; and using the trained active probability to compute the probability that the document has changed. Applicants respectfully submit that the subject matter of Claim 3, particularly when taken in combination with the subject matter of Claims 1 and 2, is clearly not taught or even remotely suggested by Chakrabarti et al. Thus, applicants submit that Claim 3 is clearly allowable for reasons in addition to the reasons why Claims 1 and 2 are allowable.

LAW OFFICES OF
CHRISTENSEN O'CONNOR JOHNSON KINDNESS^{LLP}
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100

Claim 4 is dependent upon Claim 3 and recites that the method further comprises selecting the probability that the document has changed from the previous crawl as the active probability in the current crawl and repeating the method of Claim 3 for the current crawl. Again, applicants respectfully submit that this subject matter is not taught or remotely suggested by Chakrabarti et al. and, thus, that Claim 4 is allowable for reasons in addition to the reasons why Claims 1-3 are allowable.

Claim 5 is dependent upon Claim 3 and recites that training the active probability includes multiplying the active probability indicative of a change in the document by training probability calculated using a probabilistic model. Again, this subject matter is not taught or remotely suggested by Chakrabarti et al. and, thus, Claim 4 is submitted to be allowable for reasons in addition to the reasons why Claims 1-3 are allowable.

Claim 6 is dependent upon Claim 1 and recites the probabilistic model further comprises: training a document probability distribution corresponding to the document address specification to reflect experience with the document during a plurality of previous crawls, the document probability distribution including a plurality of probabilities; determining from the document probability distribution a probability that the document has changed; and making a determination of whether to access the document in a current crawl based on the probability that the document has changed. This subject matter is also not taught or remotely suggested by Chakrabarti et al. Thus, Claim 6 is submitted to be allowable for reasons in addition to the reasons why Claim 1 is allowable.

Claim 7 is dependent upon Claim 6 and recites that the method further comprises: calculating, based on the experience of the document during a plurality of previous crawls, a discrete random variable distribution that includes a plurality of training probabilities; and multiplying each probability in the document probability distribution by a corresponding training probability from the discrete random variable distribution. As with the other dependent claims, this subject matter is not even remotely suggested by Chakrabarti et al. and, thus, Claim 7 is submitted to be allowable for reasons in addition to the reasons why Claims 1 and 6 are allowable.

Claim 8 is dependent on Claim 7 and recites that training probabilities are calculated using a Poisson process, the Poisson process including a particular Poisson equation and a complementary Poisson equation. Again, this subject matter is not taught or even remotely suggested by Chakrabarti et al. and, thus, Claim 8 is submitted to be also allowable for this additional reason.

LAW OFFICES OF
CHRISTENSEN O'CONNOR JOHNSON KINDNESS^{PC}
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100

Claim 9 is dependent upon Claim 8 and recites that the experience with the document during the plurality of previous crawls is derived from historical information associated with the document address specification. Since Chakrabarti et al. does not teach anything with respect to historical information, again, clearly the subject matter of Claim 9, particularly when considered in combination with the subject matter of the claims from which Claim 9 depends, is not taught or remotely suggested by Chakrabarti et al. Thus, Claim 9 is submitted to be allowable for this reason as well.

In summary, dependent Claims 2-9 include additional recitations that further distinguish the claimed subject matter from the teachings of Chakrabarti et al. As a result, applicants respectfully submit that these claims are allowable for reasons in addition to the reasons why the claim from which these claims directly or indirectly depend (Claim 1) is allowable.

As amended, independent Claim 10 reads as follows:

10. A computer-readable medium having computer-executable instructions for retrieving one document in a plurality of documents from a remote server, which when executed comprise:

maintaining historical information associated with changes to the one document;

initiating a crawl procedure for retrieving particular documents in the plurality of documents; and

determining whether to access the one document from the remote server based on a probabilistic analysis of the historical information associated with the changes to the one document.

Clearly, Chakrabarti et al. teaches nothing whatsoever regarding determining whether to access a document from a remote server based on a probabilistic analysis of historical information associated with changes to the document. Because Chakrabarti et al. clearly does not anticipate Claim 10, applicants submit that Claim 10 is allowable.

Applicants also submit that all the claims that depend from Claim 10, i.e., Claims 11-20, are allowable for the same reasons that Claim 10 is allowable.

Applicants further submit that Claims 11-20 are allowable for additional reasons. Claim 11 recites that the instructions further comprise: If the determination to access the document (which is based on historical information per Claim 1) is positive, identifying the one document for retrieval during the crawl procedure; and attempting to retrieve all documents identified for retrieval during the crawl procedure. Claim 12, which depends from Claim 10,

LAW OFFICES OF
CHRISTENSEN O'CONNOR JOHNSON KINDNESS^{LLC}
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100

recites that the probabilistic analysis comprises computing a probability that the one document is changed since the one document was last retrieved from the remote server. As noted above, Chakrabarti et al. does not teach any form of probability computation.

Claim 13, which depends from Claim 12, recites that computing the probability that one document has changed further comprises beginning with a probability that a predefined portion of the documents in the plurality of documents has changed, training the probability that the predefined portion of documents has changed using historical information associated with the one document to achieve the probability that the one document has changed. Again, this subject matter is not taught or even remotely suggested by Chakrabarti et al.

Claim 14, which depends from Claim 12, recites that the instructions further comprise making a random decision to retrieve the one document wherein the random decision is based on the probability that the one document has changed. Claim 15 is dependent upon Claim 14 and recites that the random decision is further biased by a synchronization level configured to influence the random decision based on a predetermined degree of tolerance for not retrieving the one document that the document is likely to have changed. Claim 16 is dependent upon Claim 14 and recites that the random decision is made by a software routine adapted to simulate the flip of a coin.

Clearly, the subject matter of Claims 11-16 is not taught or suggested by Chakrabarti et al., particularly when the subject matter of these claims is considered in combination with the subject matter of the claims from which they depend. Thus, applicants submit that these claims are allowable for reasons in addition to the reasons why Claim 10 is allowable.

Claim 17 is dependent upon Claim 10 and recites that: the historical information associated with changes to the one document includes a time stamp for the one document, the time stamp being indicative of the time that the one document was last modified when the one document was last retrieved from the remote server; and the probabilistic analysis includes a comparison with the time stamp included in historical information with another time stamp associated with the one document stored on the remote server. Claim 18 is dependent upon Claim 17 and recites that if the time stamp included in the historical information does not match the other time stamp associated with the one document stored on the remote server, identifying the one document for retrieval during the crawl procedure. Again, applicants respectfully submit that the subject matter of Claims 17 and 18 is not even remotely suggested by Chakrabarti et al. and, thus, these claims are allowable for reasons in addition to the reasons why Claim 10 is allowable.

LAW OFFICES OF
CHRISTENSEN O'CONNOR JOHNSON KINDNESS^{LLC}
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100

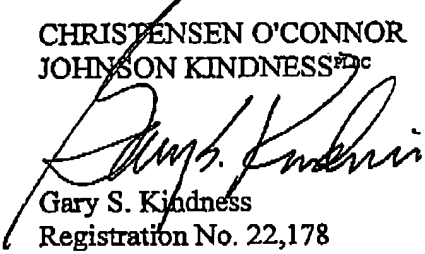
Claim 19 is dependent upon Claim 10 and recites that the historical information associated with changes to the one document includes a hash value associated with the one document, the hash value being a representation of the one document; and the probabilistic analysis includes a comparison of the hash value included in the historical information with another hash value calculated from information retrieved from the one document on the remote server. Claim 20 is dependent upon Claim 19 and recites that if the hash value included in the historical information does not match the other hash value associated with the one document stored in the remote server, identifying the one document for retrieval during the crawl procedure. Again, the subject matter of Claims 19 and 20, particularly when considered in combination with the subject matter of Claim 10, is clearly not taught or even remotely suggested by Chakrabarti et al. Thus, Claims 19 and 20 are also submitted to be allowable for reasons in addition to the reasons why Claim 10 is allowable.

CONCLUSION

In view of the foregoing amendments and comments, applicants respectfully submit that all of the claims in this application are clearly not anticipated by Chakrabarti et al. and, thus, are allowable. Consequently, early and favorable action allowing these claims and passing this application to issue is respectfully solicited. If the Examiner has any questions, the Examiner is invited to contact applicants' attorney at the number set forth below.

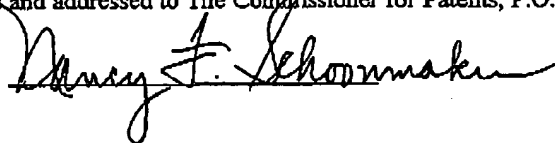
Respectfully submitted,

CHRISTENSEN O'CONNOR
JOHNSON KINDNESS^{PC}


Gary S. Kindness
Registration No. 22,178
Direct Dial No. 206.695.1702

I hereby certify that this correspondence is being deposited with the U.S. Postal Service in a sealed envelope as first class mail with postage thereon fully prepaid and addressed to The Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on the below date.

Date: 2/12/04



GSK-jch/nfs

LAW OFFICES OF
CHRISTENSEN O'CONNOR JOHNSON KINDNESS^{PC}
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100